L’uso del sistema Nuovo Soggettario per l’indicizzazione semantica di risorse web: problemi e proposte

Elisa Bianchi, Maria Clotilde Camboni, ElenaLazzarini

1 Il contesto

L’obiettivo di questo contributo è illustrare alcune questioni e spunti di riflessione relativi all’indicizzazione semantica di risorse web svolta nell’ambito del progetto ”Panoramafirb”.[1] Il progetto Panora-

[1]Il progetto Panoramafirb (FIRB RBNE07C4R9), http://www.panoramafirb.it, è finanziato dal Ministero dell’Istruzione, dell’Università e della Ricerca (Decreto n. 190/Ric., 12 marzo 2009), e si concluderà nel giugno 2013. Le autrici del presente contributo hanno collaborato alla definizione della struttura e alla costruzione del catalogo dei siti web Panoramafirb descritto nei paragrafi seguenti. Partecipano al progetto i Dipartimenti di Studi Italianistici, Storia delle arti e Informatica dell’Università di Pisa, il Dipartimento di Italianistica e spettacolo dell’Università degli studi di Roma ’La Sapienza’, il Consorzio ICoN - Italian Culture on the Net, la Direzione Generale per i Beni Librari e gli Istituti Culturali del Ministero per i beni e le attività culturali e Cap s.p.a. Nell’ambito del progetto, è stata avviata una collaborazione con il Sistema Bibliotecario d’Ateneo (SBA) dell’Università di Pisa e, attraverso di esso, con la Biblioteca Nazionale Centrale di Firenze (d’ora in poi, BNCF). Per l’aiuto preziosissimo che hanno fornito, desideriamo ringraziare le bibliotecarie del Sistema Bibliotecario d’Ateneo Cinzia Bucchioni, Francesca Cecconi, Anna Colotto, Daniela Fiaschi, Anna Delogu, Chiara Garzetti, Maria Picciani, Cinzia Romagnoli, Elisabetta Soldati, Paola Spinesi, Simona Turbanti.

JLIS.it. Vol. 4, n. 2 (Luglio/July 2013).
DOI: 10.4403/jlis.it-8828

mafirb ha preso le mosse dalla constatazione che è molto difficile, soprattutto per un utente non esperto, reperire in rete risorse di qualità relative alla lingua e linguistica, letteratura e arte italiane.[2] Il web infatti si presenta oggi come un enorme repertorio non organizzato di siti, pagine e materiali di varia natura, in cui contenuti rilevanti e qualitativamente validi sono mescolati a tanti altri estremamente scadenti. Ai fattori qualitativi e quantitativi si aggiunge la dimensione temporale, lungo la quale le risorse web possono essere ordinate secondo un progressivo grado di permanenza vs. instabilità dei contenuti, in alcuni casi intrinsecamente legato al tipo di sito (i contenuti di un blog, per esempio, saranno molto meno stabili di quelli di un sito istituzionale o di una rivista online).[3] L'obiettivo principale del progetto era favorire l'orientamento dell'utente "non esperto" nella selezione e valutazione di risorse in rete relative a lingua e linguistica, letteratura e arte italiana, attraverso due strumenti: un catalogo di siti e un metamotore di ricerca sul web. In questo contributo desideriamo impostare una riflessione sull'esperienza (in parte ancora in corso) di indicizzazione per soggetto delle risorse web attraverso

[2]Sull'argomento non esistono studi sistematici ma, limitatamente al dominio di lingua e linguistica italiana, un quadro sintetico ma esaustivo dei contenuti disponibili è fornito dai due contributi di Mirko Tavosanis pubblicati nella sezione "Lingua italiana" del Magazine online di Treccani: "La lunga marcia attraverso il web" http://www.treccani.it/magazine/lingua_italiana/speciali/divulgazione/Tavosanis.html e "L'italiano (e la grammatica) nel web" http://www.treccani.it/magazine/lingua_italiana/speciali/grammatica/Tavosanis.html.

[3]La questione della persistenza, autorevolezza e affidabilità dei contenuti web e dei problemi di catalogazione ad essa connessi è già stata trattata da Lunghi et al, che ne discutono in relazione al passaggio dai documenti ai linked data, ed è una questione molto complessa, cui in questa sede è opportuno solo accennare. Esiste inoltre una consistente bibliografia sulla definizione di standard catalografici per le risorse elettroniche, tra cui segnaliamo le due corpose monografie di Stefano Gambari e Mauro Guerrini (*Definire e catalogare le risorse elettroniche*; *Le risorse elettroniche. Definizione, selezione e catalogazione*).

il sistema Nuovo Soggettario,[4] svolta nell'ambito dell'allestimento del catalogo: in particolare, ci proponiamo di tracciare un quadro dei termini del Thesaurus NS usati, dei nuovi termini che abbiamo proposto di inserire e delle criticità e spunti di riflessione emersi nella selezione dei contenuti da indicizzare e nella costruzione delle stringhe di soggetto.

Il catalogo, che allo stato attuale consiste di circa 1000 schede, adotta con alcune modifiche (v. infra, § 2) il modello di dati del progetto europeo MICHAEL (Multilingual Inventory of Cultural Heritage in Europe[5]) e prevede tre tipi di schede: Istituzione, Servizio e Collezione. L'indicizzazione semantica dei contenuti web riguarda le due schede Servizio e Collezione, e ad essa è dedicato il campo "Tema", dove si trovano una o più stringhe di soggetto costruite in base al metodo pre-coordinato. Ciascuna stringa corrisponde a un'unità di contenuto del sito o della collezione catalogata, intendendo con "unità di contenuto" l'intero sito, se le informazioni e tipi di dati ivi contenuti sono caratterizzati da una certa omogeneità, oppure una sottosezione di esso, o ancora un insieme di dati o materiali che possano essere contrastivamente descritti come un insieme distinto rispetto agli altri contenuti presenti (v. § 2).

2 Il modello di dati

Ai fini della costruzione del catalogo, è stato adottato il MICHAEL Data Model, ma è stato necessario apportarvi alcune modifiche, che lo rendessero più adatto a descrivere le risorse di interesse per il progetto. Infatti nel MICHAEL Data Model il principale oggetto di interesse per l'utente, e quindi del catalogatore, è la "collezione

[4] http://thes.bncf.firenze.sbn.it e Biblioteca Nazionale Centrale di Firenze.
[5] http://www.michael-culture.org.

digitale",[6] ma solo una parte dei contenuti online da censire ai fini del progetto Panoramafirb poteva essere considerata tale (al massimo tra un quarto e un quinto delle risorse oggetto di catalogazione). Inoltre, mentre il progetto MICHAEL si propone principalmente di censire il patrimonio culturale digitale, non di darvi accesso (tanto è vero che sono catalogati DVD e altre risorse accessibili solo in loco), uno degli scopi principali di Panoramafirb era permettere all'utente finale l'immediato reperimento delle risorse in rete, fornendogli l'indirizzo preciso (URL con link) del sito web in cui era possibile rintracciare i contenuti desiderati. La necessità di associare le risorse catalogate ad un indirizzo web era inoltre legata a vincoli tecnici connaturati alle esigenze di sviluppo del progetto (in particolare ai meccanismi di indicizzazione del metamotore di ricerca sviluppato nell'ambito del progetto stesso), che determinavano anche i requisiti per la scelta di una certa URL tra le diverse che spesso rinviano alla stessa risorsa. L'oggetto "risorsa online", su cui doveva focalizzarsi la catalogazione, finiva quindi con l'avvicinarsi molto al "sito web".

Nel MICHAEL Data Model, i siti web sono solo uno dei possibili servizi che possono dare accesso a una collezione, e di conseguenza sono catalogati seguendo un modello non adatto agli scopi di Panoramafirb. I problemi più rilevanti scaturivano dal fatto che nel MICHAEL Data Model le schede dei servizi non hanno una sezione dedicata al tema. È evidente che ciò, nel catalogo Panoramafirb, avrebbe pregiudicato la possibilità degli utenti di rintracciare in base ai contenuti di loro interesse la maggior parte delle risorse catalogate.

D'altro canto, il "sito web" come oggetto di catalogazione pone problemi almeno in parte già noti.[7] Il concetto viene correntemente

[6] Va qui osservato che le collezioni digitali di MICHAEL sono molto spesso l'esito della digitalizzazione di collezioni tradizionali.

[7] Tali problemi vengono affrontati nel quadro dei diversi progetti che si propongono di studiare soluzioni per l'archiviazione del Web, tra i quali si può citare in

usato e compreso senza apparenti difficoltà, ma non ha una definizione formale univoca. Intuitivamente, si può dire che un sito web è un insieme di pagine Internet che si trovano nello stesso dominio web: perlopiù è vero, ma non sempre.

Alcuni siti infatti occupano più di un dominio: ad esempio, il sito della rivista online Bollettino '900[8] permette di leggere il numero in corso nel dominio web http://www3.unibo.it, ma dalla ricerca nel sito si arriva agli articoli dello stesso numero tramite il dominio http://www.boll900.it (la situazione è ancor più complessa, ma in questa sede è inutile approfondire). Molto più spesso accade che all'interno dello stesso dominio si trovino contenuti assai eterogenei, più o meno distinguibili: http://www.maldura.unipd.it/italianistica/ALI ospita una bibliografia sulle autrici italiane dei sec. XIX-XX; http://www.maldura.unipd.it/ami/php corrisponde all'Archivio metrico italiano (database di versi con marcatura degli accenti metrici, da opere dei secoli dal XIII al XVI); http://www.maldura.unipd.it/masters/italianoL2/Lingua_nostra_e_oltre rimanda a una rivista che si occupa di aspetti teorici e applicativi dell'apprendimento e insegnamento dell'italiano come lingua seconda; http://www.maldura.unipd.it/alci è una pagina dove vengono riuniti i lemmari di più repertori di neologismi.[9]

Vista questa situazione, è stato inevitabile adottare una soluzione di compromesso che permettesse ai catalogatori il massimo della flessibilità. Sono state riprese le entità Collezione e Servizio del

---

particolare Archives de l'Internet della Bibliothèque Nationale de France, soprattutto perché ha sperimentato un approccio incentrato sul sito web e non sulla singola pagina (Abiteboul et al. 7). Sul concetto di sito web si interrogano anche gli storici della rete (Brügger). Vi sono anche studi tesi a trovare un metodo per identificare automaticamente le pagine appartenenti ad un dato sito web, ma ricadono ovviamente al di fuori dell'ambito di questo contributo.

[8] http://www.boll900.it.

[9] Esiste anche un problema legato ai cosidetti mirror sites: nella nostra prospettiva aveva però un'incidenza minore rispetto ai precedenti.

MICHAEL Data Model, ma con forti modifiche, soprattutto nel caso della seconda. Rispetto al progetto MICHAEL, il focus della catalogazione e conseguentemente dell'indicizzazione semantica si è infatti spostato verso l'entità Servizio (senza dubbio quella usata più di frequente dai catalogatori), non più considerata un semplice punto di accesso ai contenuti, ma un contenitore degli stessi, più o meno coincidente col sito web. Nel modello adottato, l'entità Servizio è stata quindi arricchita di diversi campi, tra i quali "Tema" (dedicato all'inserimento delle stringhe di soggetto), "Periodo storico", eccetera. Va notato che, visto l'ambito e gli scopi specifici del progetto, sia i record Collezione che i record Servizio sono stati inoltre dotati di un campo obbligatorio "Dominio tematico", in cui segnalare se la risorsa online catalogata era pertinente all'arte, la letteratura o la lingua italiana (con la possibilità di selezionare uno, due o tutti i domini). Per la delimitazione degli oggetti da catalogare – ovvero i siti web – è stato scartato un approccio di tipo strettamente informatico (come sopra visto, pressoché impossibile) a favore di uno che, pur tenendo in considerazione il più possibile le esigenze tecniche del progetto, le armonizzasse con quelle della catalogazione e quindi con l'usabilità del catalogo da parte degli utenti. In particolare, una delle soluzioni adottate per ridurre il problema dell'eterogeneità dei materiali all'interno di una risorsa è stata la catalogazione separata di parti rilevanti chiaramente individuabili e dedicate a un argomento specifico di alcuni siti web, nei casi in cui ciò venisse considerato utile. Le schede delle sezioni catalogate separatamente sono state collegate a quelle dei siti di cui fanno parte tramite le relazioni "è parte di" e "contiene", la cui applicazione all'entità Servizio è un'altra novità del modello adottato rispetto al MICHAEL Data Model. Un esempio di applicazione di questa soluzione è il sito Italica,[10] che è stato catalogato con schede diverse corrispondenti all'intero sito

[10]http://www.italica.rai.it.

e alle sezioni di esso dedicate a Dante,[11] alla narrativa italiana del Novecento,[12] alla storia della lingua italiana,[13] al Rinascimento,[14] eccetera.

3 La scelta del sistema NS

La scelta di utilizzare il sistema NS per l'indicizzazione semantica delle risorse web è stata il punto di arrivo di un percorso di valutazione di altre risorse lessicali disponibili: abbiamo infatti preso in considerazione, in particolare, Ital-Wordnet[15] e DMOZ.[16] Ital-Wordnet è un database semantico-lessicale organizzato secondo tassonomie e relazioni lessicali codificate, liberamente consultabile in rete. Non è una risorsa lessicale disciplinare, ma contiene parole dell'italiano generale. DMOZ è un progetto di classificazione manuale di siti e risorse web attraverso l'attribuzione di etichette relative al tipo di sito, al contenuto ecc.; le etichette ("categorie") sono organizzate in tassonomie, che possono essere navigate per livelli successivi di complessità. All'interno di DMOZ, quindi, le etichette si riferiscono all'intero sito, e determinano la collocazione del sito nelle liste di risorse recensite. Il requisito centrale delle tassonomie di DMOZ è la natura "user friendly".

Sulla base delle proprietà dei due strumenti sopra illustrati, per l'indicizzazione semantica delle risorse catalogate il sistema NS è stato scelto in virtù delle seguenti considerazioni:

[11] http://www.italica.rai.it/monografie/dante.
[12] http://www.italica.rai.it/monografie/grandi_narratori_900.
[13] http://www.italica.rai.it/monografie/storia_lingua_italiana.
[14] http://www.italica.rai.it/monografie/rinascimento.
[15] http://www.ilc.cnr.it/iwndb/iwndb_php.
[16] http://www.dmoz.org/World/Italiano.

1. il thesaurus NS è un vocabolario controllato, che ha una vasta copertura disciplinare e una stretta connessione con fonti bibliografiche e lessicografiche autorevoli;

2. è uno strumento recente (la prima versione è stata rilasciata nel 2006), in continuo aggiornamento;

3. esiste un vivo dibattito terminologico tra i gruppi che ne curano l'implementazione, secondo un percorso strutturato di proposta - discussione - approvazione - validazione di nuovi termini;

4. nel corso del progetto, si è presentata l'opportunità di inserirsi nel dibattito proponendo nuovi termini (v. infra, §§ 6 e 7);

5. la combinazione dei termini nelle stringhe di soggetto, secondo la sintassi pre-coordinata, consente di descrivere una vasta gamma di contenuti con un numero limitato di termini, e quindi di produrre, nel catalogo, raggruppamenti omogenei di siti;

6. l'indicizzazione semantica attraverso le stringhe di soggetto consente all'utente di fare ricerche sia per termini singoli (analogamente alle categorie di DMOZ) sia per combinazione di termini, e quindi di consultare il catalogo delle schede Servizio e Collezione per livelli successivi di specificità e raggruppando i siti per omogeneità di contenuto con gradi diversi di granularità.

4 Problemi di scrittura delle stringhe a copertura di oggetti eterogenei e livello di granularità dell'indicizzazione semantica

Così come il modello di dati, è stato necessario adattare alle caratteristiche degli oggetti catalogati (come definiti nel § 2) anche il Sistema NS: siamo stati costretti a reinterpretare e ricontestualizzare l'obiettivo delle stringhe "coestese con il contenuto di soggetto che debbono rappresentare" (Biblioteca nazionale centrale di Firenze 101-105). Malgrado l'adozione dell'approccio di descrizione di un sito su più livelli (descritto nel § 2), non è stato infatti sempre possibile stabilire una relazione biunivoca tra stringa di soggetto e contenuti oggetto di catalogazione. Nella costruzione delle nostre stringhe è stato quindi necessario tener conto della natura miscellanea e spesso non uniforme delle risorse da catalogare. In questi casi si è ritenuto di dover utilizzare più stringhe, assimilando di fatto i siti alla tipologia (prevista nel Sistema NS) degli studi miscellanei o "Scritti in onore", in modo da rendere quanto più possibile ragione dell'articolazione dei contenuti (Biblioteca nazionale centrale di Firenze). Un caso esemplare a questo proposito è la sezione dedicata al Rinascimento[17] del sito Italica.[18] Malgrado si fosse optato per una catalogazione a più livelli, infatti, indicizzare tale sezione con il solo termine "Rinascimento" sarebbe stato insufficiente, e d'altra parte al suo interno si trovano materiali di natura eterogenea: brani di opere dell'epoca, studi di ambito rinascimentale (nelle due sottosezioni Saggi e Monografie) e altri documenti. Nel campo "Tema" della

[17] http://www.italica.rai.it/monografie/rinascimento.
[18] http://www.italica.rai.it.

scheda dedicata al sito sono state quindi inserite più stringhe, tra le quali Rinascimento – Studi e Rinascimento – Opere – Antologie.

La scelta del livello di granularità dell'indicizzazione semantica è stata comunque fatta in un'ottica contrastiva rispetto alla totalità delle risorse catalogate. La scheda del ricchissimo sito web del centro di studi dedicato a Primo Levi[19] ha nel campo "Tema" unicamente "Levi, Primo", e lo stesso accade per molti altri siti incentrati su autori riguardo ai quali si trovano risorse specifiche solo in un paio di domini web, indipendentemente dalla qualità e ricchezza (a volte notevole) dei materiali in essi presenti. Invece, nel caso delle decine di siti che si occupano di Dante Alighieri e della sua opera, molto spesso nel campo "tema" di una singola scheda sono state inserite più stringhe, per segnalare esattamente quali risorse venissero rese disponibili dal sito catalogato (testi di opere, traduzioni degli stessi, bibliografie, trascrizioni o riproduzioni di manoscritti, studi, riviste scientifiche dedicate...). La scelta di procedere in questo modo è stata compiuta pensando alle difficoltà a cui sarebbe andato incontro un utente che dopo una richiesta apparentemente specifica si sarebbe trovato di fronte a decine di risultati: in questo modo, gli sono state offerte le possibilità da un lato di sfruttare l'indicizzazione semantica per raffinare la ricerca e trovare più agevolmente i materiali a cui poteva essere interessato, e dall'altro di farsi un'idea dei contenuti dei siti web anche per mezzo delle stringhe di soggetto ad essi associate, oltre che della loro descrizione.

[19] http://www.primolevi.it.

5 Uso del ruolo "forma" (intellettuale/bibliografica)

Una delle peculiarità delle risorse web o comunque digitali rispetto a quelle librarie è la potenziale disponibilità di diverse modalità di fruizione di uno stesso contenuto. Un esempio chiaro può essere il testo di un'opera letteraria, o di più opere nel caso di banche dati o biblioteche digitali: esso può essere reso fruibile in una forma in cui può essere solamente letto (come le riproduzioni in formato immagine di testi a stampa), oppure in una modalità che permette di fare al suo interno delle ricerche. In quest'ultimo caso, si può trattare di semplici ricerche di stringhe all'interno di un unico documento, con o senza caratteri jolly, o di ricerche molto più complesse (su tutti i testi di un determinato archivio o su un determinato sottoinsieme, con ricerca delle co-occorrenze, etc.). A volte ai testi sono associati metadati anche molto raffinati, che aprono possibilità di interrogazione altrimenti impossibili: ad esempio, nel caso di DanteSearch,[20] i testi delle opere di Dante sono corredati di lemmatizzazione e marcatura grammaticale e sintattica. I testi o i risultati della ricerca possono inoltre essere scaricabili o non esserlo. Discorsi in parte analoghi e in parte diversi possono essere fatti per le bibliografie, e a contenuti diversi da questi possono essere applicate modalità di fruizione ancor differenti (ad esempio, il testo di un dizionario o di un'enciclopedia può essere offerto in modalità parzialmente o totalmente ipertestuale, nel caso in cui alcuni o tutti i termini presenti in una voce permettano di arrivare direttamente alla voce ad essi corrispondente).

È ovvio che, per un utente, le forme assunte dai diversi contenuti e le loro modalità di fruizione rivestono un interesse notevole. Di conseguenza, si è cercato di permettere la ricerca di determinate ri-

[20] http://dante.di.unipi.it:8080/DanteWeb.

sorse sulla base della loro forma e delle operazioni che vi si possono compiere, e ciò ha portato ad un notevole uso, nella catalogazione, di termini nel ruolo di forma intellettuale/bibliografica, a volte associati in serie (ad esempio Dizionari – Ipertesti o Commenti – Archivi di dati). Malgrado ciò, in diversi casi non è stato possibile fornire un'indicizzazione semantica che rendesse conto delle effettive particolarità di una risorsa. Una delle cause che concorrono a questa insufficienza è ovviamente il fatto che la terminologia del Thesaurus NS non è stata pensata allo scopo di descrivere risorse come quelle del web e in particolare del web 2.0, soprattutto quelle interattive (vedi oltre il caso di "Forum"). Va inoltre aggiunto il fatto che la riflessione e la discussione su questi aspetti sono ancora carenti, ed è ovviamente difficile risolvere questioni così complesse in poco tempo e affrontandole a partire da una prospettiva in fondo limitata. D'altra parte, la ristrettezza dell'ambito di applicazione della catalogazione unita all'indicazione della forma bibliografica/intellettuale hanno permesso di risolvere la difficile questione della catalogazione dei siti web che contengono biblioteche digitali onnicomprensive, come Googlebooks o Archive.org, e anche di quelli che mettono integralmente a disposizione le pubblicazioni scientifiche legate a una determinata università (repository di ricerca o siti di case editrici universitarie). In entrambi i casi, si è optato per indicare i soggetti di interesse per il progetto (arte, letteratura e lingua italiana) seguiti dalla forma "Biblioteche digitali".

6 Termini nuovi e combinazione di termini già esistenti

Per via delle esigenze esposte nel paragrafo precedente, nell'ambito della discussione con BNCF le proposte di nuovi termini[21] hanno riguardato soprattutto – oltre ovviamente a etichette specifiche dei domini disciplinari arte, letteratura e lingua o linguistica italiana – termini che potessero rendere ragione delle diverse forme e modalità in cui le risorse web possono essere messe a disposizione dell'utente.

Nel thesaurus si trovano alcuni termini utili per descrivere la "natura" dei siti, sia dal punto di vista tecnico che della struttura dei dati: "Weblog" per i "Blog" (forma, quest'ultima, indicata come non preferita); "Biblioteche digitali" per i siti che contengono banche dati di opere digitalizzate, "Archivi di dati" per i database (online e scaricabili). Alcuni di questi termini sono stati inseriti su proposta delle catalogatrici del progetto Panoramafirb: tra essi, "periodici elettronici" e "forum".

Per quel che riguarda "forum", è facile intuire che si tratta di un termine fondamentale per la catalogazione di molti siti, essendo attivi allo stato attuale molti forum con una tradizione ormai consolidata negli anni, che costituiscono un punto di riferimento per gli utenti della rete (per esempio, per quelli interessati alla lingua italiana il forum SoloItaliano di Wordreference[22]).

Quanto invece a "Periodici elettronici" (inserito come NT di "Pubblicazioni elettroniche"), all'interno del catalogo Panoramafirb sarebbe stato utile affiancargli "Periodici scientifici elettronici". L'u-

[21] Nel corso della collaborazione con il Sistema Bibliotecario di Ateneo dell'Università di Pisa e con BNCF, sono stati proposti e accettati in totale 9 termini "comuni" (Abstract, Collane editoriali, Edizioni elettroniche, Excerpta, Forum, Frontespizi, Italianistica, Periodici elettronici, Bollettini elettronici) e circa 20 termini disciplinari appartenenti a letteratura, lingua e linguistica e arte italiana.

[22] http://forum.wordreference.com/forumdisplay.php?f=51.

so dei due termini consentirebbe infatti di distinguere le riviste online in generale dalle riviste scientifiche, che seguono precisi percorsi e standard di revisione, indicizzazione e pubblicazione dei contributi. Tuttavia, nell'ottica di limitare il più possibile la proliferazione di termini nel Thesaurus a favore piuttosto dell'espressione di un determinato concetto tramite la combinazione di termini già presenti, BNCF ha optato per non accettare il termine.

Quello della proposta di "Periodici scientifici elettronici" è stato uno dei tanti casi in cui è sorta una questione di grande rilievo, che concerne i meccanismi di equilibrio tra due esigenze contrapposte ed egualmente importanti: quella di utilizzare un vocabolario controllato ed esprimere concetti differenti attraverso la combinazione dei termini nella sintassi delle stringhe di soggetto da una parte e quella di "anticipare" le query dell'utente con termini di uso comune dall'altra.

Così, un termine che abbiamo proposto è stato "Edizioni elettroniche", che abbiamo definito come "Edizioni pubblicate in formato elettronico, destinate alla lettura e a funzioni avanzate di ricerca e di elaborazione dei contenuti." Questo termine sarebbe molto utile per descrivere le edizioni (nel senso filologico del termine) create in formato digitale e raccogliere in un unico gruppo omogeneo i numerosi siti frutto di progetti che avevano come scopo la creazione di edizioni digitali di opere: ad esempio le opere di Dante lemmatizzate, o ancora le grammatiche digitalizzate (come immagini e come testo) della Biblioteca dell'Accademia della Crusca.

Nel dibattito che è stato avviato su questo termine, BNCF ha proposto di scomporre "Edizione elettronica" in "Edizioni, Pubblicazioni elettroniche", tanto è vero che, allo stato attuale, Edizione elettronica è indicato come termine non preferito, e rimanda all'uso dei due termini combinati.

La continua dialettica tra la necessità di avere un vocabolario

controllato e uniforme da una parte e di rappresentare in maniera unitaria la specificità (terminologica e concettuale) di un settore disciplinare, dall'altra, si fa ancora più stringente nel caso della descrizione di risorse web per un catalogo elettronico, in cui l'utente accede al catalogo attraverso query, di cui il catalogatore deve in qualche modo tenere conto: così, un termine centrale per le ricerche sul web, "E-learning", nel thesaurus BNCF è indicato come forma non preferita, da sostituire con la combinazione di tre termini, cioè "Educazione, Impiego, Internet": tale divergenza tra indice di frequenza (e verosimilmente, di familiarità) di un termine nelle ricerche sul web e controllo del vocabolario nel thesaurus pone un interessante spunto di riflessione (allo stato attuale, una questione aperta) sul rapporto tra gli standard di indicizzazione semantica e le query digitate dagli utenti.

7 Termini di dominio utilizzati e termini nuovi

7.1 Arte italiana

Per quel che riguarda i termini o le categorie di termini utilizzati nell'indicizzazione semantica di risorse online relative al dominio Arte Italiana è stato necessario ricorrere a elementi non presenti nel Thesaurus, come i nomi propri di artisti e di opere. Tra i termini inseriti su nostra proposta si può citare "Pittura ferrarese". Per quanto riguarda le forme bibliografiche/intellettuali, nelle schede di Arte ricorrono frequentemente i seguenti termini: Collezioni, Monumenti, Collezioni digitali.

7.2 Letteratura italiana

Per quel che riguarda i termini o le categorie di termini del Thesaurus del NS più adoperati nell'indicizzazione semantica di risorse online relative alla letteratura italiana, un ruolo molto importante è stato giocato – com'era d'altra parte prevedibile – da elementi nel Thesaurus non presenti, ovvero i nomi propri (in particolare di autori e opere). Il termine in assoluto usato con la massima frequenza – e anche questo era prevedibile – è "Letteratura italiana"; altri termini specifici (come "Letteratura drammatica italiana", "Letteratura dialettale sarda", "Poesia per musica", etc.) sono stati adoperati molto più di rado. La "categoria di appartenenza" dei termini che nel loro insieme sono adoperati più spesso è però quella delle diverse etichette atte a ricoprire il ruolo di forma intellettuale/bibliografica, come: Opere, Edizioni, Libretti, Manoscritti, Descrizioni, Riproduzioni, Studi, Testi, Ipertesti, Biografie, Traduzioni, Periodici, Indici, Autografi, Incunaboli eccetera. Ciò appare una conseguenza scontata di quanto detto nel § 5; altra meno scontata è il fatto che i nuovi termini proposti e accettati per l'inserimento nel Thesaurus sono in buona parte termini non specificamente riconducibili alla letteratura italiana ma necessari per indicizzare in maniera adeguata alcune risorse: Abstract, Collane editoriali, Edizioni interpretative, Forum, Frontespizi, Periodici elettronici, Periodici umbri. Nella maggior parte dei casi, le nuove proposte riconducibili alla letteratura italiana rientrano nella serie "Luoghi carducciani", "Luoghi folenghiani", "Luoghi leopardiani"... Il termine di maggior rilievo disciplinare tra quelli entrati nel Thesaurus su nostra proposta è "Italianistica", che però non è specifico dell'ambito letterario, dal momento che la disciplina copre anche gli studi rivolti alla lingua italiana.

7.3 Linguistica italiana

Per quanto riguarda il dominio di linguistica italiana, i termini del thesaurus NS (e le combinazioni di termini) maggiormente usati rappresentano le due principali aree tematiche in cui possono essere raggruppati i siti web dedicati alla lingua e linguistica italiana:

Lingua italiana - Insegnamento [agli] Stranieri
Lingua italiana – Grammatica

Per quanto riguarda le forme bibliografiche/intellettuali, nelle schede di linguistica ricorrono frequentemente i seguenti termini: Biblioteche digitali, Corpora, Forum (termine inserito da BNCF su nostra proposta) e Weblog (forma preferita di "Blog"). Tali termini rappresentano molto bene la grande bipartizione dei siti dedicati alla lingua e linguistica italiana: da una parte, infatti, abbiamo siti "divulgativi", creati da e per utenti non specialisti e che hanno un interesse generico per la lingua italiana (forum, blog); dall'altra, una parte consistente di siti catalogati appartiene all'ambito scientifico/accademico, ed è costituita da biblioteche digitali e corpora testuali destinati alla comunità scientifica. Una questione aperta, che riteniamo opportuno attualizzare in questa sede, è rappresentata dalla catalogazione di materiali didattici di varia natura e pubblicati in vari tipi di siti: si tratta di documenti scaricabili (per esempio in formato .doc o .pdf), di pagine html o intere sezioni di siti contenenti esercizi, progettazione di percorsi didattici, spunti per attività di vario genere, indicazioni bibliografiche destinate ai docenti di lingua italiana sia come lingua materna che come lingua seconda.Ora, a prescindere al problema dell'autorevolezza e persistenza di questi documenti, abbiamo comunque ritenuto opportuno catalogare e indicizzare queste risorse con stringhe di soggetto dedicate, utilizzando gli unici termini disponibili nel thesaurus: Schede didattiche

ed Esercizi. Sarebbe senz'altro proficuo, non solamente per il dominio disciplinare della lingua e linguistica italiana, articolare meglio la tassonomia dell'etichetta di nodo [Parti e proprietà di documenti], introducendo termini specifici per documenti e siti del web progettati per scopi didattici (sia per docenti che per studenti).

8 Conclusioni

L'esperienza di costruzione del catalogo Panoramafirb dei siti web ha permesso di approfondire la questione della definizione di "sito web", "risorsa elettronica" e in generale della descrizione e catalogazione dei contenuti web. Inoltre, l'indicizzazione semantica delle risorse catalogate ha evidenziato la necessità di adattare gli strumenti terminologici esistenti (nel nostro caso, il NS) attraverso due percorsi: in primo luogo, la creazione di termini nuovi sia disciplinari (di lingua, linguistica e arte italiana) sia comuni (relativi alla natura dei contenuti descritti), in secondo luogo l'elaborazione di criteri specifici per la segmentazione dei contenuti e la descrizione attraverso le stringhe di soggetto. Il risultato è un catalogo uniforme dal punto di vista della soggettazione, in cui a stringhe di soggetto simili corrispondono risorse affini, e in cui quindi è possibile ravvisare una corrispondenza biunivoca tra contenuto e termini/stringhe di soggetto. Certamente, la costruzione del catalogo ha posto alcune questioni che rimangono tuttora aperte e che richiederebbero di essere sviluppate nell'ambito di ricerche ulteriori: in particolare, riteniamo che le due linee di sviluppo più promettenti siano da una parte i criteri di identificazione (e catalogazione) univoca delle risorse web selezionate, dall'altra lo sviluppo di funzionalità del database del catalogo che massimizzino il potenziale esplicativo e di raggruppamento di risorse omogenee delle stringhe di soggetto e dei termini del thesaurus NS.

Riferimenti bibliografici

Abiteboul, Serge, et al. «A First Experience in Archiving the French Web». *Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries*. ECDL '02. London: Springer-Verlag, 2002. 1–15. (Cit. a p. 5).

Biblioteca nazionale centrale di Firenze. *Nuovo soggettario. Guida al sistema italiano di indicizzazione per soggetto. Prototipo del Thesaurus*. Milano: Bibliografica, 2006. (Cit. a p. 9).

Brügger, Niels. «L'historiographie de sites Web: quelques enjeux fondamentaux». *Le Temps des Médias* 18.1 (2012): 159–169. (Cit. a p. 5).

Gambari, Stefano e Mauro Guerrini. *Definire e catalogare le risorse elettroniche*. Milano: Bibliografica, 2002. (Cit. a p. 2).

——.*Le risorse elettroniche. Definizione, selezione e catalogazione*. Milano: Bibliografica, 2002. (Cit. a p. 2).

ELISA BIANCHI, Consorzio ICoN.
e.bianchi@italicon.it

MARIA CLOTILDE CAMBONI, Università di Pisa.
m.c.camboni@humnet.unipi.it

ELENA LAZZARINI, Università di Pisa.
e.lazzarini@arte.unipi.it

Bianchi, E., M. C. Camboni. A., A. Lazzarini. "L'uso del sistema Nuovo Soggettario per l'indicizzazione semantica di risorse web: problemi e proposte". *JLIS.it*. Vol. 4, n. 2 (Luglio/July 2013): Art: #8828. DOI: 10.4403/jlis.it-8828. Web.

ABSTRACT: The essay deals with the creation of subject-headings for web resources related to Italian literature, art and linguistics with resort to the Nuovo soggettario system (NS). It describes the difficulties arisen and the results achieved in this regard during the development of a project aimed at creating web tools to facilitate the location of high quality web resources about Italian culture. One of these tools was a catalogue of web resources with subject headings, created using a modified version of the MICHAEL Data Model. The authors explain why they had to change the

model to meet the needs set by the peculiar items of the catalogue, why the choose the NS for the subject-headings, their choices about the granularity of the description, their particular use of the "Intellectual/Bibliographic form" roles of the NS to match the features of their items that could be relevant for a user, and their consequent proposals of new terms for the NS Thesaurus and the questions that arose from these proposals.

KEYWORDS: Digital cataloguing; Semantic indexing; Nuovo Soggettario

ACKNOWLEDGMENT: L'articolo è frutto di un lavoro condiviso; Elisa Bianchi ha scritto i §§ 1, 3, 6, 7.3 e 8; Maria Clotilde Camboni ha scritto i §§ 2, 4 5 e 7.2; Elena Lazzarini ha scritto il § 7.1.

Submission: 2013-03-14
Accettazione: 2013-04-08
Pubblicazione: 2013-07-01